DRIVER: Building a sustainable infrastructure for global repositories

Dale PETERS1, Norbert LOSSAU2

1DRIVER Scientific Technical Manager, Göttingen State and University Library, Platz der Göttinger Sieben 1, 37077, Göttingen, Germany
Tel: +49 551 39 5242, Fax: +49 551 39 5222, Email: peters@sub.uni-goettingen.de

2DRIVER Scientific Coordinator; Director Göttingen State and University Library, Platz der Göttinger Sieben 1, 37077, Göttingen, Tel: +49 551 39 5212, Fax: +49 551 39 5222 r, Email: Lossau@sub.uni-goettingen.de

Abstract: DRIVER is a bold vision. The project represents the vision of worldwide networks of scientific data repositories and aims to achieve a global infrastructure supporting the scholarly communication of the future.

The primary objective of DRIVER is to establish a flexible, robust, and scalable infrastructure for all European and world-wide digital repositories, managing scientific information in an Open Access model increasingly demanded by researchers, funding organisations and other stakeholders. It continues to expand its content base with high quality research output, including textual research papers, data sets and other scholarly publications. The recent D-NET v1.0 open source software release forms the basis of this distributed service-oriented architecture that enables enhanced interoperability of data and service-providers, with functionality ranging from search, recommendation, collections, profiling to innovative tools for repository managers. By building a robust network of content providers, enhanced with the set of services and software tools that DRIVER offers, the DRIVER infrastructure also enables further collaboration to domain specific research communities in a co-ordinated network of repositories from a growing number of institutional repositories, from national research institutions and data aggregators.

Exciting new developments will see the extension of the infrastructure to global research communities in China, India, Africa and Latin America, deploying a vigorous awareness and advocacy programme. The DRIVER Confederation offers a support service for repository managers, a dynamic set of guidelines aimed at data interoperability, and the strategic support for new forms of scholarly communication.

This paper will address DRIVER’s response to the global information environment, in concepts of scientific and technical collaboration that currently both challenge repository managers, and promise a future to open access publication through the growing international repository network. These issues are addressed in terms of recent technical developments, but include also strategic issues of policy development, improved services and additional functionality offered to the scholarly community that point to the achievement of the ultimate goal of a global and interactive representation of human knowledge.

Keywords: Open Access; Institutional repositories; Scholarly communication.
1. Introduction

DRIVER has a bold vision: *All research institutions in Europe and worldwide make all their research publications openly accessible through institutional repositories.* The vision follows the Berlin Declaration, which called in October 2003 for “free and unrestricted access to sciences and human knowledge representation worldwide”. Initiated by the internationally renowned German research organisation, the Max-Planck-Society, and signed by many international research organisations and institutes, the Berlin Declaration has set a political statement. In building a sustainable infrastructure for scientific repositories, DRIVER brings to this statement the reality of scholarly communication in the future.

In response to the Berlin Declaration, the DRIVER initiative set out in 2005 as a multinational European Union test-bed project under the FP7 e-Infrastructures programme. Designed to explore the development of a distributed infrastructure to enable enhanced interoperability of data, DRIVER is currently into its second phase, having achieved a network of digital repositories containing research and other scholarly publications from twelve partner institutions or national clusters across ten European countries.

The main deliverables are currently to provide a production quality digital repository infrastructure and to establish a European digital repository Confederation. These objectives will, in conjunction, achieve a cross-country community of content providers and an infrastructure for the international networking of open access repositories.

A nucleus and early inspiration for DRIVER was the national DARE¹ programme in The Netherlands, which released a network and services for institutional repositories in 2004. One of the identified success factors of the DARE programme was community building. The affiliation in the DRIVER Confederation is based similarly, on a willingness to comply with a set of principles and some agreements related to organisation, data and technology.

In response to an increasing demand for open access to scientific information, the current imperative lies in extending the DRIVER infrastructure to the global research community. This paper will address the envisaged role of the DRIVER Confederation to extend the flexible, robust, and scalable infrastructure to all European and world-wide digital repositories. The methods applied in the internationalisation of the DRIVER Information Space will be discussed, to include repositories and add content from countries within and beyond Europe. A secondary objective in achieving a global infrastructure is framed by the challenge of innovative scholarly communication, to include not only institutional repositories, but also data collections of subject repositories and learned societies. Thirdly, new elements of the infrastructure and new services will be discussed as mechanisms to facilitate the global expansion programme.

1. Objectives

The internationalisation programme aims to power the geographical expansion and deepen the data infrastructure of the repository ecosystem to stimulate e-Infrastructure initiatives
based on regional, national and international networks of repositories. This is achieved in the following objectives:

1. To afford significant economies of scale in providing the necessary expertise, knowledge and experience to ‘new entrants’ to the repository scene.
2. To apply the support mechanisms, the organisational structures, management protocols and policy guidelines developed by the wider research community in building a large-scale and interoperable international scientific data infrastructure.
3. To add value in critical mass and in a wide variety of multimedia formats to the existing base of scientific content

2. Methodology

From the outset, DRIVER has adopted a result-driven approach to structure the consortium activities. In considering the experiences learned from the DARE programme, DRIVER identified four main areas of activity:

1. Organise digital repository content providers and build a cross-country community as a Confederation;
2. Develop an open source software suite to network physically distributed repositories;
3. Implement this software as operational instance of a digital repository infrastructure;
4. Provide an online DRIVER Portal

3.1 Community and Confederation

The building of an informed repository community as a prerequisite for a functioning infrastructure. A key lesson of the DARE programme in The Netherlands, and confirmed by other advanced national communities like SHERPA in the UK, has been the relevance of active community building. The nature of such communities is varied, DRIVER sees communities in various contexts: national, subject-based, technical (e.g. repository platform providers like ePrints, D-Space or Fedora).

The rationale for building an informed repository community is based on multiple factors. For example, communication is more efficient between twenty national communities than between five hundred local institutional repositories. Technical and data quality requirements aimed at local repositories are more effectively discussed and disseminated in the constraints of the local environment. A significant problem for OAI service providers like OAIster, BASE or DARE.Net lies in the effort required to manually correct or “clean” records submitted by individual repositories. Jackson (et al) stress that despite the importance of disseminating descriptive metadata capable of supporting interoperability, the manner in which institutions are implementing Dublin Core in practice is seldom discussed. Improving the quality of metadata (e.g. by standardising usage of fields like DC:author [Name], [Surname]) remains an enormous challenge to interoperability and harmonising the application profiles for local OAI-PMH implementations would significantly support automated harvesting processes.

Community building also involves both the strategic level (mainly library directors) and the operating level (repository manager and team) to support the network mission and the necessary implementations. This is especially important where additional effort is required to improve data retrospectively. Communities are further characterised by a certain level of commitment from their members. Again, this commitment is a prerequisite for a
reliable, large-scale infrastructure in the use, for example, of persistent identifiers and the guarantee for long-term access to online publications in repositories.

A central objective of the current internationalisation programme therefore, is to establish a Confederation of digital repository communities. As organisational backbone of the DRIVER infrastructure it is envisaged beyond the duration of project funding, to ensure the sustainability of both the infrastructure and the organisational network of repository communities.

3.2 Internationalisation of the DRIVER Information Space

The expansion of the Confederation provides an appropriate mechanism to advance the adoption of OA policy among data providers. The challenge of innovative scholarly communication, widens the definition of data providers to include not only institutional repositories, emerging in academic and research libraries, but also data collections of subject repositories and learned societies, representing the research community. Representative candidates in EU-funded projects, like CLARIN (Common Language Resources and Technology Infrastructure Network); DARIAH (Digital Research Infrastructure for the Arts and Humanities) and NEEO (Network of European Economists Online). The research-generating members of the European Science Foundation (ESF) such as the Max Planck Institute are likely candidates to lead the endorsement of self archiving of publically-funded research.

Expressions of interest received from China, India, Africa and Latin America, suggest that a planned response is required to meet the growing demand in a dedicated programme. The provision and maintenance of DRIVER Information Space is European-based, but the data infrastructure has to be global in its coverage of international content, for it to be of optimal use to science and research. Certain research areas, such as health, climatology and geomorphology can no longer be adequately served regionally. Worldwide there are many efforts on a national, regional, local, community and/or consortium scale, but none of the scope and ambition that is characterized in DRIVER. The level of progress also differs significantly on different continents. Some countries already count various local, regional or consortium (networks of) repositories. In other countries, activities focus on awareness building and advocacy, by organising conferences and other information events, but have not evolved to building up an infrastructure yet. Open Access concepts have been accepted in Asian countries, but they still lack super-national infrastructures.

The increasing importance of countries and parts of the world that have previously been marginalised in the global research arena makes it necessary to include their research output – as much as needing access to the global knowledge economy as a means of building a just social agenda.

3. Technology Description

3.1 Software development

In June 2008 the DRIVER Consortium released version 1.0 of the DRIVER Network Evolution Toolkit (DNET)” under an Open Source Apache License. The license type was chosen after discussions that included commercial software companies, in order to support widest possible collaboration.
DNET Version 1.0 offers 23 services in total, and supports three groups of users:

- The repository network administrator is supported to harvest, clean, enrich and aggregate data from local repositories (e.g. by the Repository Network Manager, Resource Monitoring)
- The local repository manager can check compatibility of local data with the DRIVER index format (DRIVER Validator)
- End-users, and in particular, researchers can use the virtual knowledge base with some fundamental features (search, browse, profiling)

From a software architecture point of view, DNET recognises three main layers:

- Enabling service management layer
- Data management layer
- End-user service functionality layer

![Figure 1 DRIVER Production Infrastructure](image)

The service management layer enables the interaction between all layers, the dynamic addition and grouping of single services and service monitoring. The services can be used in combination or stand-alone. Increasingly, new services are being added by DRIVER network partners or third-party service providers. The D-NET v. 1.0 software release is also made available to any organization willing to run independent installations of the DRIVER software; and able to assume responsibility for the maintenance of such infrastructure. With this release, DRIVER demonstrates its support for the development of further open infrastructures; and its willingness to promote other service providers and relative repository communities.

4.2 Infrastructure Implementation
A possible deployment scenario of the DNET software is envisaged in the implementation by a repository community to set up a national repository network. Such an example is already evident in RECOLECTA, the Spanish national service for the search, access and retrieval of full text open access scholarly publications of Spanish academic and research institutions. A joint collaboration project between REBIUN (Network of Academic and Research Libraries of Spain) and FECYT (Spanish Foundation for Science and Technology), the aims of RECOLECTA are twofold:

- Coordinate and stimulate the development of a network of interoperable digital scientific repositories as a mean of a national knowledge infrastructure based on the open access principles and open and international standards and protocols.

- To develop, and allow third parties to develop, a set of services on top of the aggregated content, targeting different needs: search, re-use, online collaborative tools, citation measurement, subject portals, national web portals, etc.

The DRIVER Infrastructure offers technology that can be customized to the specific needs of such organizations willing to build a uniform Information Space from an arbitrary number of heterogeneous OAI-PMH data sources. Communities may thus define applications operating over the infrastructure that are specific to their needs and that can be implemented as a customization and extension of the DRIVER Infrastructure.

D-NET has been developed as a technically advanced, open source toolkit for re-use by repository networks. Deployment thereof guarantees interoperability, saves costs and time, but is not mandatory. More critical to the functionality of a shared infrastructure is the use of DRIVER services and data models to support enhanced interoperability in a distributed infrastructure. The infrastructure software enables the objectives of achieving an open access network of digital repositories containing research and other scholarly publications across the globe.

4. Developments

4.1 DRIVER Portal Development

In the interim it has become clear that networking repositories means more than harvesting metadata through the OAI protocol. Existing service providers report many problems arising from heterogeneous implementation of this protocol. Interoperability at the price of manual correction of data is to be avoided, as this is not scalable. At the same time workload for local repository providers needs to be contained within acceptable limits.

The online portal service offers basic usability in search, browse, personal profiling and support for research communities. The service functionality relies on repository community support by DRIVER, allowing repository managers to share and exchange knowledge and experience around initial setup of IR’s and the organizational aspects of IR management. In addition, the portal service relies on raising professional awareness in areas of repository data management:

- High quality of data is a priority, as it is the basis for automated services like profiling, statistics, metrics, and also for links to research evaluation systems (CRIS);
• Delivery of open access full text;
• Interoperability is achieved through the common set of metadata policies, as described in the DRIVER guidelines, and offers maximum exposure of research data.

The DRIVER Guidelines, now in a second version, are maintained by a volunteer metadata expert group comprising DRIVER partners and other international experts to improve the quality of available metadata and to assist managers of new repositories in defining their local data management policies. The harmonisation of OAI-implementations and application profiles serve too, in avoiding harvesting errors, ultimately ensuring a degree of uniformity across all repositories in the global information platform.

5. Results

The inclusion of individual repositories into the DRIVER infrastructure is possible, but national communities are preferred as natural partners in the global repository network. One institution or a group of institutions are encouraged to take responsibility for building a national repository community like DARENet in the Netherlands; SHERPA in the UK; OA.Net in Germany; RECOLECTA in Spain; and HAL in France and DRF in Japan. Experience has shown that national communities are best represented by country correspondents, designated to liaise with the DRIVER community, reporting national status of repository development, organising local events, and translating repository guidelines into local languages. This effective and dynamic interaction is maintained in a wiki format, available online, to identify local contact details to encourage membership and support of an existing national community. National nodes may in time build up national data aggregators, cleaning data and providing additional services to the local community as required.

6. Business Benefits

Where no national nodes exist for repository aggregation as outlined above, dedicated repository managers are encouraged to indicate on the DRIVER wiki their interest in taking up this role, and thus joining the DRIVER Confederation. The benefits of joining the Confederation are many fold:

• By joining the DRIVER community individual repositories contribute to the successful implementation of Open Access principles in a powerful, international community.
• Certain research areas, such as health, climatology and geomorphology can no longer be adequately served regionally. Local researchers require an international infrastructure in which their research output becomes visible as part of a common international knowledge base.
• Local funding bodies and research organisations can build their own interface on the DRIVER data index to demonstrate national aggregations that meet their own needs.
• National communities become new data providers, enjoying the transferral of technological expertise, guidelines, support, promotion and advocacy services in promoting the uptake and usage of local repositories.
• National communities share in technical developments contributed to the DRIVER Infrastructure by service providers.
The DRIVER Confederation, in which the members assume responsibility for the DRIVER objectives, contributes data, offers services, shares expertise and suggests strategic direction to the repository community. These objectives form strong natural incentives, beyond project funding, to build an international repository community that is firmly embedded in national communities, and supported by significant alliances with international partners.

7. Conclusions

DRIVER’s mission is to expand its content base with high quality Open Access research output, as well as provide support for repository managers and state-of-the art services for the end-user. By deploying the DRIVER Confederation to extend the flexible and scalable infrastructure to all European and world-wide digital repositories in a robust network of content providers, enhanced with the complex set of services that DRIVER offers, the DRIVER Infrastructure also enables service providers to develop new applications on top of the basic services. It acts as a showcase for repository development, as a networking tool for the DRIVER community and demonstrates a range of end-user services. This community is defined in the DRIVER Confederation, comprising data providers; service providers; the major repository software platforms; representatives of emerging standards, protocols and data models; and NGO’s serving the repository community.

However, much remains to be done. The lessons that have been learned from the DRIVER project point to a number of recommendations:

1. The need to balance and strengthen technical services and software achieved in infrastructure development, with increased advocacy for Open Access and scholarly communication. The coordinated approach presented in the DRIVER Confederation will undoubtedly facilitate the implementation of an international networking of Open Access repositories.

2. The lobby for Open Access mandates from all funding organisations is high on the list of priorities for the repository community - if we are to achieve a global research infrastructure that successfully attracts the interest and serves the needs of the scholarly community. As an alternative reward system to that offered by current journal impact factors, new bibliometrics are needed to measure usage and citations of documents in Open Access repositories, in order to secure that essential quality assurance upheld by the principle of peer review.

3. Greater effort is required in building Open Access consortia to revise contracts with journal publishers, to convert current subscription payments to parallel publishing models, i.e. “subscription + institution-wide ‘green road’ IR deposit permission”. The pilot Springer Open Choice agreements are indicative of potential future developments in this area. Another example is seen in the role of DRIVER in co-ordinating repository interaction with publishers in the PEERxiv project. A pioneering collaboration between publishers, repositories and the research community, PEER (Publishing and the Ecology of European Research) will investigate the effects of the large-scale, systematic self-archiving on reader access, author visibility, and journal viability, as well as on the broader ecology of European research.

4. Co-ordinated action is required to support both Open Access journal publishing and Open Access book publishing, as seen in the European OAPEN project.xv

5. Digital curation and long-term preservation services are urgently required by the repository community, if we are to meet the challenge of curating enhanced publications, comprising research articles and related datasets.
DRIVER provides a voice to national repository communities; makes them visible in the European and international context. The development of local institutional repositories is accelerated, with strong economies of scale achieved in DRIVER’s shared services, which can be built on top of local repositories. Such a network replicates and innovates traditional modes of scholarly communication to allow comparative searching within subject domains - and can help to fill repositories, as researchers, research institutes and universities want to have maximum visibility - best realised in a network of content repositories.

Five years ago, the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities acknowledged the chance to constitute a global and interactive representation of human knowledge, including cultural heritage and the guarantee of worldwide access. DRIVER “Digital Repository Infrastructure Vision for European Research”, has come to represent this vision of world-wide networks of content repositories, offering a robust infrastructure supporting scholarly communications of the future.

References


iii (2008). Jackson, A.S, Han, M-J, Groetsch, K., Mustafoff, M., Cole, T. Dublin Core Metadata harvested through OAI-PMH. Journal of Library Metadata. 8(1) p.5-21


v RECOLECTA: http://www.recolecta.net/buscador/


vii DARENet is the network of Digital Academic Repositories in the Neetherlands. It is a result of the national DARE programme that aimed to coordinate and stimulate the development of repositories containing scientific output in the Netherlands. Since April 2008 DAREnet has been integrated in the scientific portal NARCIS. http://www.narcis.info/index/tab/darenet/

viii TheSHERPA partnership has within its membership a range of examples of repository environments and institutional structures within research-led institutions, offering the ideal environment for exploring and testing ideas for repository development, which can be evaluated and disseminated to the wider community. http://www.sherpa.ac.uk/index.html

ix Supported by the DFG, the project OA-Network is a joint collaboration of the Humboldt Universität zu Berlin and the Universities of Göttingen and Osnabrück. It aims to virtually integrate all document and publication services with a DINI certificate and to increase the number of DINI certified repositories. http://www.dini.de/projekte/oa-netzwerk/

x RECOLECTA: http://www.recolecta.net/buscador/
The HAL archive is a national aggregator which offers a common platform to multiple archives comprising French universities, major higher education schools and major research institutes which have signed a common protocol to make the HAL archive system their common platform to host the national research output. [http://www.driver-support.eu/national/france.html](http://www.driver-support.eu/national/france.html)


DRIVER wiki information on national communities: [http://www.driver-support.eu/national/index.html](http://www.driver-support.eu/national/index.html)

PEER Project to Examine Large-Scale "Green" Open Access Impacts. [http://digital-scholarship.org/digitalkoans/2008/10/13/peer-project-to-examine-large-scale-green-open-access-impacts/](http://digital-scholarship.org/digitalkoans/2008/10/13/peer-project-to-examine-large-scale-green-open-access-impacts/)

Open Access Publishing in European Networks (OAPEN) is a project in Open Access publishing for humanities and social sciences monographs. The Open Access movement has developed rapidly in the sciences and in journal publishing. The consortium of University-based academic publishers who make up OAPEN believe that the time is ripe to fully explore the possibilities of Open Access for the humanities and social sciences. [http://www.oapen.org/](http://www.oapen.org/)